**Data Driven Wine Selection**

*Selection Massale*

COMP 4441: Probability and Statistics 1

Logan Hahn & Chidi Nzerem

## **Research Question**

Our research question; Can we predict Wine Quality at Quinta Da Aveleda using physicochemical characteristic data? Researchers in the Vinho Verde region gathered physicochemical data on all the white wines in a winery called Quinta da Aveleda. In the spirit of *Selection Massale,* our team will use linear regression with interactions and quadratic variables to predict the scores of a holdout test set of unscored wines, skim the top 200 from that selection which could be fed into an optimization model that would narrow the selection down to a 6-12 wine showcase that maximizes predicted score.

## **Data Understanding**

The data used for this assignment downloaded from OpenML, came from a set of researchers in the Vinho Verde region, who collected data on the main components of white wines as well as their overall quality. For this project, we will be using the dataset to predict the rating of a wine before its release to be certified based on physicochemical characteristics recorded during production. The dataset had 11 different numeric input features that are based on the physicochemical tests, and a factor output variable of Quality (Wine Certification) that is determined of a score between 1-10 with the highest score being seven and the lowest score being one.

The input variables consisted of different types of acidity, sugars, sulphates, and alcohol acting as different features in predicting the quality of a wine.

The data collected by researchers in the Vinho Verde region is described by the following characteristics:

* Shape: (4898 rows, 12 columns)
* All columns are continuous numeric variables
* Column distributions below

|  |  |
| --- | --- |
| **Fixed Acidity -** *Most acids involved with wine or fixed or nonvolatile (do not evaporate readily)* |  |
| **Volatile Acidity -** *The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste* |  |
| **Citric Acid -** *Found in small quantities, citric acid can add ‘freshness’ and flavor to wines* |  |
| **Residual Sugar -** *The amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet* |  |
| **Chlorides -**  *The amount of salt in the wine* | \*Chloride values between .01-.35 |
| **Free Sulfur Dioxide -** *The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine* |  |
| **Total Sulfur Dioxide: -**  *Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine* |  |
| **Density -**  *The density of water is close to that of water depending on the percent alcohol and sugar content* | \*Density Values between .99 and 1.04 |
| **pH -** *Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale* |  |
| **Sulphates -** *A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant* |  |
| **Alcohol: % Alcohol by Volume (ABV) -** *The percent alcohol content of the wine* |  |
| **Quality:** *Represents the certification rating of a wine based on sensory and physicochemical characteristics. Higher is better.* |  |

## **Regression Model**

### **Preparation**

While the dataset that was provided byresearchers in the Vinho Verde region was relatively clean, we found that there were a few things that we had to go through and tweak to get the data ready for analysis. When downloading the data directly from Google OpenML, the column names were encoded V1 through V11 with the Quality section being renamed to Class, so the columns had to be changed to their actual names. To deal with the encoding of the features in the dataset, we used a package called “tidyverse” to replace the V encoding to the actual column names. This dataset contained no null values. The only other cleaning that was needed after changing the column names and checking for null values was to change the Quality column to a numeric value from a factor value to allow the model to accurately read and spit out numerical values.

### **Model Development**

In the initial modeling phase, we tried several different predictive modeling tools before arriving at our final model using a linear regression. First, we created a decision tree models, subsequently creating a logistic regression model. We found that the tree models did not provide the optimal results we were hoping for. The Class variable, converted to “Quality” in our training dataset, is a categorical variable ranging in labels from one to seven, therefore, when we looked at using a binomial logistic regression, we had to transform Quality into a binary variable in order to use it as our target variable. We hoped to use the logistic regression model to identify specific wines that met the average quality criteria of the batch of wine we were using in our analysis. After transforming the data and running a logistic regression, we decided the most efficient model to use would be a linear regression. Regarding the linear regression, there were certain benefits of using this modeling technique, primarily being, we could predict the relationship between the continuous Quality variable along with various explanatory variables. Ultimately, the best predictive model that we used to analyze this data was a linear regression.

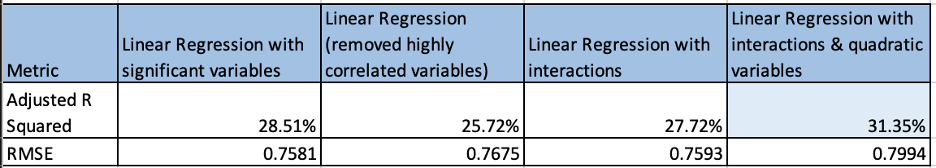
### **Final Model: Linear Regression With Interactions & Quadratic Variables**

The model was used to predict the Quality rating of a test set of wines explained by the independent variables included in the model. We trained and tested our model with various interactions between explanatory variables as well as adding data transformations until we found our best fitting linear model.

The first step was checking the multicollinearity of the variables. This was important because we didn't want our independent variables that were highly correlated to each other causing instability in our regression. When checking for multicollinearity, we set our variance inflation factor to a threshold maximum of 10, leading us to remove “residual\_sugar” and “density” as predictors. We also looked at the correlation of each independent variable against the target variable of Quality, using individual scatterplots. Through our analysis and interpretation of the scatterplots, we found that many of the variables were negatively correlated with the Quality of the wine. These negatively correlated variables include fixed acidity, volatile acidity, residual sugar, chlorides, total sulfur dioxide, and density. The analysis showed that generally, as these variables decreased, the Quality rating of the wine increased.

Some important information found in the regression analysis was that the interaction between volatile\_acidity and alcohol had the highest effect on the Quality of wine with a coefficient of 7.076e-01. Meaning that for every increase in Quality there was a 7.076e-01 increase in the interaction between volatile\_acidity and alcohol. A close second was the interaction between fixed\_acidity and alcohol. Although when there was interaction between alcohol, fixed\_acidity and volatile\_acidity there was a positive correlation to the Quality. Standing alone, three features have a negative correlation to Quality.

In addition to including interaction terms in our model, we also transformed some of the variables and added quadratic terms. The quadratic terms were added to improve the overall fit of the model. When observing the linearity of each variable, there was evidence of some of the variables not having a fully linear relationship with the dependent variable. In order to model the effect of the variables more accurately and find the best fitting model, we created a quadratic variable for alcohol, fixed\_acidity, and free\_sulfur\_dioxide. When looking at the quadratic effect of other variables, there was no significant effect on the model. We also looked at using the log transformation to improve the model fit but did not find it necessary as none of the included variables had a high skew.



### **Evaluation**

To evaluate the model we looked at the Multiple R-squared value, 0.3155, and the Adjusted R-Squared value, 0.3135. This means that 31 percent of the variation in Quality can be explained by our model’s independent variables. We were also able to calculate the confidence interval for our coefficients. All of which we could say with 95 percent confidence that we were close to or had the true slope. Finally, after running our chosen best model on the validation dataset and running the predictions, we were able to calculate the root mean square error of 0.7994. RMSE is a good measure of how accurately the model predicts the response, and is one of the most important criterias for fit if the main purpose of the model is prediction. The RMSE continually increased in our tested models after removing the variables with high collinearity. Despite the chosen model having the highest RMSE out of all of our tested models, it is still the best fitting model because it is not undermining the statistical significance of the independent variables and has the strongest relationship with the dependent variable. Future regression modeling could include random forest regression. Lastly as we can see from the residual plot, which is the error between a predicted value and the observed actual value. We can see that the errors are independent and normally distributed.

### **Closing Statement**

Based on the work done we were able to predict with 79% accuracy the quality of wine. For further deployment, a small group of 200 wines from our predicted group can be imported to excel. This will allow the analyst to then perform further optimization to pick the top wines for various circumstances. This can be done by applying constraints in excel and will give you a more specific look at what wines are best for that specific purpose.

## **Sources:**

V. (n.d.). Portuguese vinho Verde White. Retrieved April 24, 2021, from <https://www.vivino.com/wine-styles/portuguese-vinho-verde-white>

Teixeira, F. (2017, August 17). Best wineries you can't miss in the vinho verde region. Retrieved April 24, 2021, from <https://blog.winetourismportugal.com/best-wineries-vinho-verde-region>

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009, June 09). Modeling wine preferences by data mining from physicochemical properties. Retrieved April 26, 2021, from <https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?casa_token=x9OffCJ_4MwAAAAA%3APShbZV5nVKtflIAloN3YsFScVuxAP8SDrQyjw4npUVLpEjULSFJCyn544GsB7YKTvArf3rmTLQ#bib11>

Vanschoren, J. (2016, July 29). Wine-quality-white. Retrieved April 26, 2021, from <https://www.openml.org/d/40498>

FAOSTAT. (2021, April 26). Retrieved April 26, 2021, from <http://www.fao.org/faostat/en/#data/TP/visualize>